A Hybrid Approach Solution for web page segmentation using Clustering and Optimizing Techniques

SHORT SYNOPSIS

Submitted in partial fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

Manav Rachna University

by

Meena Chaudhary
2K16-PFE-012

Under the supervision of

Dr. Jyoti Pruthi
Professor

Department of Computer Science and Technology
School of Engineering

MANAV RACHNA
Vidyaprataribha

MANAV RACHNA UNIVERSITY
Declared as State Private University vide Haryana Act 26 of 2014

Year *2022*

MANAV RACHNA
MANAV RACHNA UNIVERSITY
Declared as State Private University vide Haryana Act 26 of 2014

Date: 05/11/2022

DECLARATION BY THE SCHOLAR

I hereby declare that the short synopsis entitled "A Hybrid Approach Solution for web page segmentation using Clustering and Optimizing Techniques" being submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in the Department of Computer Science and Technology under the School of Engineering of Manav Rachna University, Faridabad during the academic year 2022

It is the work that is expected to be carried out under the supervision of Dr. Jyoti Pruthi Further, it is declared that the short synopsis is within the plagiarism limit and copy of the same is attached as per the University norms.

Meena Chaudhary

(Meena Chaudhary)
2K16-PFE-012

MANAV RACHNA
MANAV RACHNA UNIVERSITY
Declared as State Private University vide Haryana Act 26 of 2014

Date: 05/11/2022

CERTIFICATE BY SUPERVISOR(S)

This is to certify that, the short synopsis entitled "A Hybrid Approach Solution for web page segmentation using Clustering and Optimizing Techniques" is an original piece of work done by research scholar Meena Chaudhary under the supervision of Dr. Jyoti Pruthi for the degree of Doctor of Philosophy in the Department of Computer Science and Technology under the School of Engineering of Manav Rachna University, Faridabad during the academic year 2022.

Further, it is certified that the short synopsis is within the plagiarism limit and copy of the same is attached as per the University norms.

Jyoti

Signature of the Supervisor
Dr. Jyoti Pruthi,
Professor
Computer Science and Technology,
Manav Rachna University, Faridabad

ABSTRACT

Data extraction from the Web is a popular research topic as of late, with most suggested methods depending on parsing HTML code or tag trees. More and more Web databases are being available online and may be queried using Web interfaces. Traditional crawler-based search engines have a hard time indexing these unique Web sites because of the way they are created. The following are the primary restrictions of these approaches: To begin with, they rely on HTML, the language used to create Web pages. However, HTML is constantly evolving. So as new versions and tags are released, older works will need to be revised accordingly. The question raised by previous answers is whether major customization or full assignment is required. Second, it cannot handle the increasingly complex HTML used on websites. Website designers employ more sophisticated implementations of JavaScript and CSS to create visually striking and engaging websites. Compare different websites based on embedded text, images, video files, hyperlinks, etc. using Vision-based Page Segmentation Algorithm (VIPS). Gram models and lemmatization capabilities for implementing deep web data extraction from unstructured sources.

The Internet and modern applications include vast amounts of material, which may be overwhelming to sort through. Organizing massive amounts of text documents into meaningful clusters, as the text clustering approach does, is a useful tool for handling such data. Larger documents have less of an impact on the efficiency of the text clustering method. Accordingly, the efficacy of the text clustering approach is diminished due to the presence of sparse and uninformative features (i.e., noisy, irrelevant, and superfluous characteristics) in the text documents themselves. Meta-heuristic algorithms have recently been used to effectively address various challenging optimization issues. This thesis highlights the aforementioned clustering issue and suggests a few meta heuristic methods for resolving it.

Key words: skip gram, VIPS - Vision based Page Segmentation,Deep Web,*Text Clustering,Segmentation*

CONTENTS

S. No.	Description	Page No.
1	Introduction	1-5
	1.1 WEB PAGE SEGMENTATION	
	1.1 .1 Page Rank	
	1.1.2 Vision Based Page Segmentation	
	1.1.3 Skip Gram -Model	
	1.2 Clustering	
	1.2.1 Basic Clustering Concepts and Need for Clustering	
	1.2.2 Clustering Algorithms	
2	Literature Review	6-11
	2.1 Vision Based Deep Web Data Extraction	
	2.2Barriers in Deep Web Mining	
	2.3 Clustering Methods in Deep Web Data Mining	
	2.4 Evolutionary Clustering Approaches in Deep Web Data Extraction	
3	Description of Broad Area	12-13
4	Objectives of the Study	14
5	Methodology to be adopted	15-17
6	Expected Outcome of the Research	18
7	References	19-23

1. Introduction

1.1 WEB PAGE SEGMENTATION

Web sites are becoming intricate since they are seldom created by hand but rather by Content Management Systems (CMS). This makes automated analysis difficult (i.e. recognising and categorising various parts from Web sites including primary content, navigation, user comments, and advertising). Web page segmentation offers a remedy to this problem. Segmentation of a website is the act of breaking it up into smaller, more manageable chunks, or "blocks," that make sense both aesthetically and semantically.

Web page block detection is an important stage in several processes, including mobile app development, information retrieval, and Web archiving. Segmentation, for instance, may be used in the context of Web archiving to isolate and save relevant portions of content. It also enables the identification of relevant modifications (changes in key blocks) across different versions of a page by assigning relative weights to blocks according to their relevance. Crawling optimization benefits from this feature since it enables the fine-tuning of crawlers to increase the frequency with which they return sites that have undergone significant modifications. By comparing the original and updated versions of a page, it aids in controlling curation processes like converting a Web archive from ARC to WARC format. If there is no difference between the before and after segmentation, then it is likely that the activity was carried out appropriately. Segmentation is used by mobile apps in order to improve the display of Web pages on tiny screens. Zooming in on a web page is one method through which mobile devices provide information to the user. Using this method takes a long time and has a slow reaction. To improve reaction time and user experience, devices may employ segmentation to display the user with relevant chunks without having to zoom in. As a means of transitioning from one format to another, web archives may use the segmentation of web pages. Web page migration from

1.1.1Page Rank

To show the reports in a requested way, Page Ranking strategies are connected, which can mastermind the records in request of their significance, significance and substance score. Web crawlers utilize two various types of positioning elements: query dependent factors and question Independent Factors .Query-subordinate are on the whole positioning variables that are particular to a given inquiry, while question free factors are joined to the archives, paying little mind to a

given inquiry. Inquiry subordinate elements utilized via web search tools are measures, for example, word archives recurrence, the position of the question terms inside the record or the modified report recurrence, which are for the most part measures that are utilized as a part of conventional Information Retrieval. A portion of the question free factors are Link prominence, Click prevalence and upto-dateness and so on. Positioning calculations in light of connection ubiquity, falls under Link based positioning calculation class.

1.1.2 Vision Based Page Segmentation

According to [Sreekanth &Narayanan ,2017], this method is the vision-based approach that recovers the semantic architecture of a progressively based page. Each hub to square map represents a different aspect of the semantic structure. The intelligibility of a square is determined by a value that is a degree of consistency and is based on the user's level of visual knowledge. The process of page division currently consists of the following three phases.

- Computerized Frame Extraction.
- Separator Identification.
- Material system building.

Website page Segmentation towards Information Extraction. Vision-Based Page Segmentation Algorithm: Extended and Perceived Success [K. Kise,2014]

This programme eliminates the drawbacks of the VIPS algorithm. VIPS has certain shortcomings due to an incomplete characterization of the edge and rule set required for visual square extraction. The Java design authors provided a new twist on the VIPS algorithm. Furthermore, VIPS does not include HTML5 attributes in the label bundle, which is a concern. In addition, the number of available labels will increase with the addition of HTML5 labels. In addition, the VIPS algorithm decides which labels to use, and they are gathered accordingly. When that time comes, the hubs created at line breaks will have been ranked according to the amount of free space they generate. Recognizable content on a page may also be identified by use of visual features such as background colour, text style colour, and text size. To understand and explain Web page material, this hybrid solution combines visual, consistent, and site page design abilities. It was suggested by the designers of the Square O-Matic system [Zhang, Jiang, and Sun, 2009] This function, used to divide a webpage, determines the direction the current is flowing. There are three steps to this structure: analysis, understanding, and replication. The

document object model tree allows you to dissect a webpage, learn its inner workings, and reconstruct the page's logical structure from scratch. The final product is a semantic tree with an integrated semantic architecture. That design's suggested model is shown in the following Fig. The document object model tree may be accessed through rendering in a web application. The algorithm of D2C evaluates the document object model tree data to get the Wcontent data structure. The output of the D2C computation using the Document object model tree as input is Wcontent. Content organisation considers what is actually shown and described on the page. Understanding archives requires converting material into a logical format. It is also managed with the help of the Web page c21 algorithm. For example, [Cormier, 2016] suggests a way for vision-based website pre-processing in which a visual engine is used to get a representation of the website, which the Rec approach then transforms into a utilitarian tree W.' It collects the smallest portions of the image using an individual computation, and then combines them into one uniform region. The bunching measures rely only on visual features, such as the size and similarity, of the individual components. The results have been shown to be comparable to the VIPS demonstration, and he has two interesting advantages over the existing algorithms. First and foremost, the technique's precise use of visual data makes it adaptable to many reports, and it can also easily adapt to new advances in HTML, DOM, and other web-based formats. The advantage that follows is the detailed and versatile production structure used for supplementary training.

1.1.3 Skip Gram -Model

- Skip Gram is used to predict the context for given a word[YesuRaju&KiranSree,2013]
- The input to the model is word si, and the output could be word si −1,wordsi −2,wordsi +1,wordsi +2,wordsi −1,wordsi −2,wordsi +1,wordsi +2.
- for each word, skip gram uses a distinct vector representation and the internal structure of words is ignored.
- Given a word w, $Gw \subset \{1, \ldots, G\}$ is the set of n-grams that appears in w and a vector which represents zg for each n-gram g.
- We represent a word by the total of the vector representation of n-grams.So the scoring function can be given by [YesuRaju&KiranSree,2013]

- $s(w, c) = \sum_{g \in G_w} Z_g^T V_c$

1.2 Clustering

Clustering is a method that has shown to be useful in many different study fields, which has contributed to its widespread adoption in the area of data mining, image processing, many applications in medical research, etc [Madhulatha,& Soni,2012.]. The rise in popularity of this algorithm can be attributed to the development of a number of new data mining algorithms over the past few years and to their widespread application in a diverse range of fields and industries. Some examples of these fields and industries include image processing, computational science, mobile communication, pharmaceutics, and financial matters. The fact that the technique for data mining cannot be institutionalized is the primary drawback associated with it [Madhulatha,& Soni,2012].

1.2.1 Basic Clustering Concepts and Need for Clustering

The purpose of clustering is to determine an unlabelled data set's natural grouping. Despite the fact that deciding how to choose is a great example of clustering. One may demonstrate that the ultimate goal of clustering cannot be determined by using the best principle. As a result, this concept must be provided by the consumer in a manner that the clustering's output is suitable to the customer's requirements [Gaertler &Marco, 2005].

1.2.3 Clustering Algorithms

Various similarity metrics are implemented as subroutines in general-purpose algorithms for clustering. The final clustering's desirable qualities should guide the algorithm selection process. Time and spatial complexity are further factors to think about [Xu, Rui,Donald Wunsch,Cai, Fan & Tahar Kechadi,2016.]. Clustering algorithms seek for similarities between data points in order to discover naturally occurring clusters of components. The technique for clustering also determines the centre of mass of a collection of data sets. Most methods use the distance between a location and the cluster centroids as a criterion for membership. The results of a clustering algorithm are often statistical descriptions of the cluster centroids and the numbers of components in each cluster. There have been several approaches to clustering put forward

throughout time. In what follows, we will go through the methods in detail [Jain, K.,Murty, &Flynn,1999].

- PARTITIONING CLUSTERING

Algorithms for dividing and clustering data typically divide a set of n items into k mutually exclusive subsets. The minimum number of objects in a cluster is one, and every object has precisely one cluster to which it belongs [Reynolds, P., Richards, Iglesia &Smith,2006]. Each data point is assigned to precisely one cluster thanks to partitional clustering's ability to divide the data into distinct and non-overlapping groups. The criteria for dividing the data into clusters and the total number of clusters are predetermined [Camila N., Razente, Marcelino, Agma JM Traina, &CTraina J,2014].

- HIERARCHICAL CLUSTERING

When applied to a dataset, Hierarchical Clustering generates a hierarchical decomposition. In a hierarchical clustering, more specific clusters might be subsets of broader ones. Hierarchical clustering may be done in two ways: agglomerative and divisive. The process of agglomerative clustering is bottom-up. every thing is its own little clump. The next thing to do is combine any two nearby clusters into one. This process of merging clusters is performed until either a full cluster tree is constructed or an endpoint is achieved. Dendrograms are trees made up of clusters, and they often show a range of sizes inside each cluster [Reynolds, P., Richards, Iglesia &Smith,2006].

If every cluster has precisely one item, or if some other termination condition is fulfilled, then the clustering process will end. The decision that an item belongs to a cluster is final in standard hierarchical clustering methods. Although this feature aids in swiftly locating clusters, the clustering quality may decrease as a result. The results of a hierarchical clustering exercise might be either fragmentary or unified. To begin, divisive clustering places all of the data into a single cluster.

2. Literature Review

2.1 VISION BASED DEEP WEB DATA EXTRACTION

[Sun, Lu& Zhou,2005]: Database queries are used to ascertain Deep Web content, and the resulting records are used to store information that may be used to produce fresh pages on the fly. Data extraction from deep Web pages is a challenging task that can be solved using a vision-based method. In order to accomplish deep web information extraction, this method is often used to the visual components of online pages. This demonstrates how the vision-based technique is evaluated by using the large database set with effective data extraction.

[Liu, Wei, Meng& Meng,2009]: The authors of article expanded the VIPS Algorithm on the Accessibility Tools Framework. However, this framework does not feature a particular method for identifying and comprehending dynamic material on web pages. Because the material that is invisible is either loaded initially as set to be invisible or is always set to be invisible, writers are required to detect both the content that is invisible and the event that converts the content that was previously invisible into visible content. After its identification, the material shouldn't be further separated into its sub-blocks; rather, it should be shown as a single visual block in its entirety. Users that visit web sites within constrained circumstances benefit significantly from the inclusion of dynamic content. However, the authors' suggested extended VIPS algorithm does not expressly concentrate on these types of enlarged information, and web pages need to be segregated as a whole. It is possible that the user needs to see just the newly expanded material. Additionally, the writers stated a desire to dynamically separate the newly included material without totally reprocessing a web page.

[Akpınar,Elgin&Yesilada,2013]: The authorsconducted research on the issue of duplicate entity identification for the purpose of deep Web data integration. They start by making an observation about the qualities in one domain, and then they make a hypothesis about whether or not the functions those attributes play are definite or depending on the domain. They then offer a comprehensive solution to handle this issue, which involves attributing weights to attributes, mapping attributes, and accomplishing training sets. In the studies, they chose two sample domains (books and computers) to examine their methodology, and the results of the experiments indicate that their methodology's accuracy is satisfactory in actual applications.

[Liu, Wei& Xiaofeng Meng.,2010]: In this article the authors developed an automated method to glean information records from search engine results pages that are produced on-the-fly. Their method is language-agnostic since it relies only on the visual elements of the response page. Because of this independence, their approach is both novel and resistant to changes in the underlying code structure. The results of this research show that the ViDE system may significantly enhance the precision of the web data extraction process. While the progress made so far is encouraging, there is still room for improvement and additional research into areas such as:

- ViDE's efficiency can be increased, ViDE is limited to processing deep Web pages with a single data region, despite the fact that many deep Web pages contain multiple data regions.
- ViDE currently obtains visual information of Web sites via using the programming APIs of IE, which is a slow operation.

2.2 BARRIERS IN DEEP WEB MINING.

[Narwal, Neetu &Sharma,2016]: The authors of article presented a system that would effectively separate useful and non-informative information, as well as extract visual blocks from web pages, categorise those blocks using machine learning algorithms, and then classify the extracted blocks. In conclusion, the method sorts out the useful material and delivers an accurate and speedier adaption of content for user perception on mobile devices. Additional study into this topic may expand this to include the use of evolutionary algorithms into the optimization of web page layout.

[Mohata.&Dhande,2015]: The authors of paper investigated the web data mining technique, which is the application of data mining techniques to extract knowledge from Web data, such as Web documents, hyperlinks between documents, usage logs of web sites, and so on. The authors of paper have studied the web data mining technique. The World Wide Web has become an integral component of many businesses, as well as the lives of everyday people and organisations in today's technologically evolved world. We have conducted research on the properties of online data due to the fact that it may be found in a wide variety of forms. Given the critical nature of extracting specific data from the internet, our research has focused on developing two efficient methods for mining large amounts of data: the first is based on the Apache Hadoop MapReduce

framework, and the second is a visualization-based method known as Visual Web Mining (VWM).

[Palekar, V. R., Ali& Meghe,2012]: The authors of article concentrated their attention on the challenge of extracting data from structured websites, namely data record extraction and data item extraction. They concluded, based on an earlier study that the visual information of Web sites may be of use to us in the process of data extraction from the Web. They looked at a huge number of deep Web sites and came up with a list of noteworthy common visual traits that may be used for deep Web data extraction. Their findings are based on their observations. On the basis of these visual characteristics, the authors developed a web programming language agnostic method to extract structured data from deep Web sites. This method has the potential to circumvent the limits that were present in earlier efforts. The use of the visual aspects of deep web pages is the primary focus of this vision-based method, which is the approach is defining characteristic.

2.3 CLUSTERING METHODS IN DEEP WEB DATA MINING

In the medical profession, as well as in psychology, biology, sociology, pattern recognition, and image processing, cluster analysis is playing an essential part in the solution of a large number of issues. The process of dividing a collection of items into distinct groups, or clusters, is referred to as clustering. This is done in such a manner that the items that belong to the same cluster have a high degree of similarity with one another, whilst the objects that belong to separate clusters share a high degree of dissimilarity with one another.

2.4 EVOLUTIONARY CLUSTERING APPROACHES IN DEEP WEB DATA EXTRACTION

[Liu, Tantan & Agrawal,2012]: The method that was proposed in uses the k-means clustering algorithm to find user frequent patterns in realistic data collected from the wide-area network. It also comes to some useful practical conclusions that can help facilitate the design of targeting and recommending applications for storing the data's on web users.

[Wu & Junjie,2012]: The authors of presented four techniques that may be used to extract content blocks from other types of material. The Content Extractor algorithm detects repeating portions by looking for instances of the same block across several web pages. The Feature Extractor algorithm is able to recognise the content block because it makes use of several very

specific capabilities. The K- Feature Extractor method uses a K-means cluster to get numerous blocks, in contrast to the Feature Extractor algorithm, which only chooses a single block. In order to differentiate the various content blocks that may be found on a web page, the L-Extractor technique utilises a combination of the Vision-based Page Segmentation Algorithm (VIPS) block partitioning algorithm and a support vector machine. On a number of different websites, the 'infodiscoverer' method was compared against other methods, including content extraction and feature extraction, and found to be superior.

[Alkhatib, Bassel,& Basheer,2019]: In this paper a technique for the informationextraction was suggested. When producing summaries from the input text, the technique of sentence clustering is the approach that is applied. The input document is pre-processed via the use of tokenization and stemming techniques. The Term-Frequency / Inverse Document Frequency (TF/IDF) formula is used to get a word score, and the sentence score is determined by summing the scores of each word that makes up the phrase and placing that total in the appropriate area. In order to get an accurate and complete definition of the text, the K-means clustering method is used. A second document with the rating of the paper's sentences has been saved. After then, the sentences are rearranged such that they are becoming longer to shorter. The score with the greatest centroid is called centroid1, whereas the score with the lowest centroid is called centroid2. A concluding summary is developed when the top k sentences from each cluster have been extracted.

[Yang, Bo, Fu, D. Sidiropoulos &Hong, 2017]: Authors introduced a new feature selection technique for Web Usage Mining that is grounded on Rough Set Theory and focuses on the first phase of the process, Data Pre-processing. The problem with Web Log Files is the quantity and variety of unnecessary data that they include. The real function selection procedure makes use of two algorithms—Fast Reduct and Variable Precision Rough Set Algorithm—to categorise the right information from the web log files. Until the aforementioned two methods are established, k-means, a clustering technique, is employed to divide up the similar patterns. So, in order to categorise the chosen attributes, the algorithms are only used on a set of similarly situated objects. Therefore, this method supplied the best means by which to dispose of irrelevant information included in web log files.

[Hailei.,2020]:]Data Mining incorporates fuzzy logic at several phases to give decisions a more human feel Because of the inherent fuzziness of human behaviour, it is very challenging to

predict such patterns using conventional, deterministic Boolean Data Mining techniques. Clustering is a method used in Data Mining to divide data into groups of related items based on their similarities. If we take an example, we tend to classify the people around us into groups like friends, relatives, colleagues, superiors, stranger etc. Anyone we encounter may belong to any number of these categories. A person's relatives and friends might both include the same individual. Again, if we consider the friend group, we will find that our relationships with the individuals that make up this circle vary in terms of how close we are to them. Crisp clustering is used in traditional Data Mining systems for similar modelling, however in this case, each individual can only belong to one group. Boolean values of 0 and 1 will be used to describe whether or not a person is a member of a certain group. Crisp clustering cannot accurately depict an individual's degree of affiliation with the group to which he belongs.

The field of Web data management is complicated by the concept of clustering. Finding replicated Web sites, identifying copyright breaches, and delivering search results in an organised manner are just some of the many uses for clustering. Both offline (independent of queries) and online (in real time) approaches to clustering exist (on the results of search queries). A lot of work has gone into mining Web traffic data and dynamically categorising search engine results. Finding comparable web sites with online approaches that use link structure and content has shown to be a success. This portion of literature surveys the current state and future directions of clustering used on the Internet. It also provides an overview of the most widely used methodology and implementations for doing so, whether it be for clustering users or sources.

[**Berkhin, Pavel.,2006]:**Using the GA, the authors describe a clustering approach that may be used to find good cluster centres in the search space. Members of a random population represent different locations in the search space. In GA, the parameters of the search space are represented as strings called chromosomes. These strings are then subjected to GA operators like crossover and mutation to generate a new generation of strings. For a certain number of generations or until a termination condition is met, selection, crossover, and mutation continue. The obtained result proves the efficiency of the GA clustering method and shows its superiority over the k-means technique.

[Wensheng, Yu, Doan & Meng, 2004]: By fusing the tabu search technique with two newly explored functional procedures—packing and releasing—the authors were able to create an effective heuristic solution for the clustering issue. This method is compared to the tabu search, the k-means algorithm, and the simulated annealing algorithm to ensure its search efficiency and efficacy. The examination of similar systems demonstrates that the created technique may be used for other kinds of information categorization and pattern recognition

3 Description of Broad Area

3.1 Information Retrieval and Information Extraction

The Internet is becoming an increasingly important information resource. The majority of the world's information is stored in unstructured text, which makes it difficult to query. Despite the fact that many websites contain long lists of pages that contain structured data, this is often the case. The term "structured data" is used to describe information that follows a predefined format The core structure, such as a relational database, is used to generate the structured pages dynamically .The value of extracting structured information from websites increases dramatically since this information may then be used as the basis for complex searches. Information integration systems, which mix the data accessible on multiple websites, have historically found structured data extraction to be the most important work and a major sub-problem . As a result, many people in the database and AI fields have been working to solve the issue of extracting data from web pages, often known as the information extraction challenge.

In addition, web pages are generated using a predefined template into which data values are inserted. Template pages with static links have been seen on the open web as well. More than that, templates may be used to generate a set of records for highlighting similar things. Therefore, many web-related applications and businesses may benefit from information extraction from template sites. In their experiments, researchers concentrated on page extraction in general. It turns out that a website's schema can be easily determined using the page-level extraction procedure. The extraction of deep web pages is aided by the identification of schemas and templates .

There are now more than 20 billion web pages on the visible internet and more than 1 trillion web pages on the invisible internet, data extraction from the World Wide Web (WWW) is a substantial operation. Most websites include a variety of unrelated elements, such as menus, images, forms, and contact details. The deep Web, also known as the covered Web or unnoticed Web, is the sum total of all the databases on the Internet. In addition, data records containing the retrieved information based on query results are discovered in web pages.

In recent years, data extraction from the internet has become more important. By sending queries to Web databases, extracting the hidden or secure web pages is a complex operation. You may

also find the returned data entries in newly produced web pages They are known as "deep web pages," and they are the inner dynamically created web pages Since websites are reliant on their underlying programming language, the challenges posed by external metrics and drawbacks are elaborated in a variety of ways. Users may often access websites at all times and conduct searches in a two-dimensional medium. Common crawler-based search engines like Google and Yahoo have trouble revealing the unique dynamically created web pages. The difficulty comes from satisfying customers who want precise answers

4 Objective of Study

I. To Implement Base VIPS for Analysis and Comparison with html parsing on html3 web pages and html5 web pages.

II. To Propose Enhanced VIPS algorithm with SKIP GRAM MODEL, Stemming and lemmatization for Vision Base Web page segmentation

III. To Propose Squirrel Search Algorithm clustering(SSAC) with K- means Clustering for Web page segmentation

IV. To Proposed Squirrel Search Fuzzy Clustering Algorithm (SSFCA) with fuzzy clustering means for Web page segmentation

V. Compare Proposed SSAC and Proposed SSFCA with Harmonic search, particle Swarm Optimization, Geneticalgorithm, K-means and Magnetic optimization algorithm and prove the effectiveness of proposed algorithms over existing ones.

5 Methodology to be adopted

5.1 Work Plan and Methodology for VIPS

The proposed system removes the aforesaid drawbacks of VIPS in following manner:

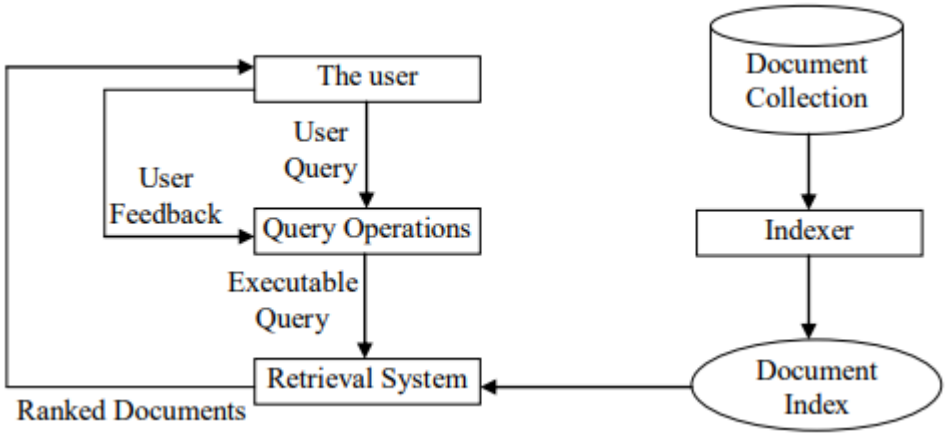

Fig 1 Proposed VIPS SYSTEM

We will start from webmining concept, we will targetany social media website like mouthshut.com for retrieval of html content with css and js files. After retrieval of unstructured data we convert it to structure data and start search for relevant words or phrase or document after applying preprocessing technique. Now our application will search relevant document on the basis of query input given by the user and display the relevant output after applying Page rank technique and Proposed VIPS algorithm with skip gram model and lemmatization function.

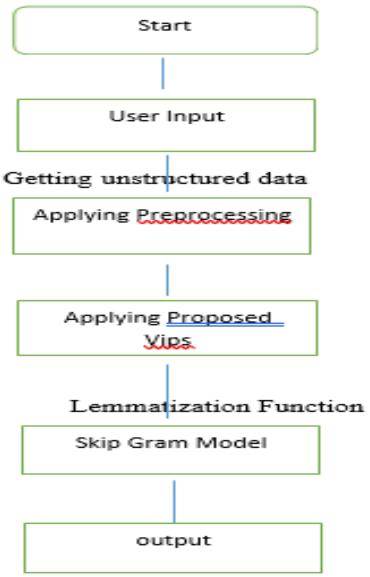

Fig 2 Technical flow of the work for proposed VIPS

5.2 Work Plan and Methodology for KMEANS and FUZZY C MEANS

A difficult task for any type of clustering algorithm is the initial selection of clustering centers. Sensitivity lies in its inability to handle outliers, noise, and imbalanced data. To overcome this problem, a squirrel search fuzzy clustering algorithm has been proposed. In this algorithm, the squirrel moves around the search space to find the best location for the centroid. To introduce the SSFCA algorithm, the process is briefly illustrated in Figure 3..

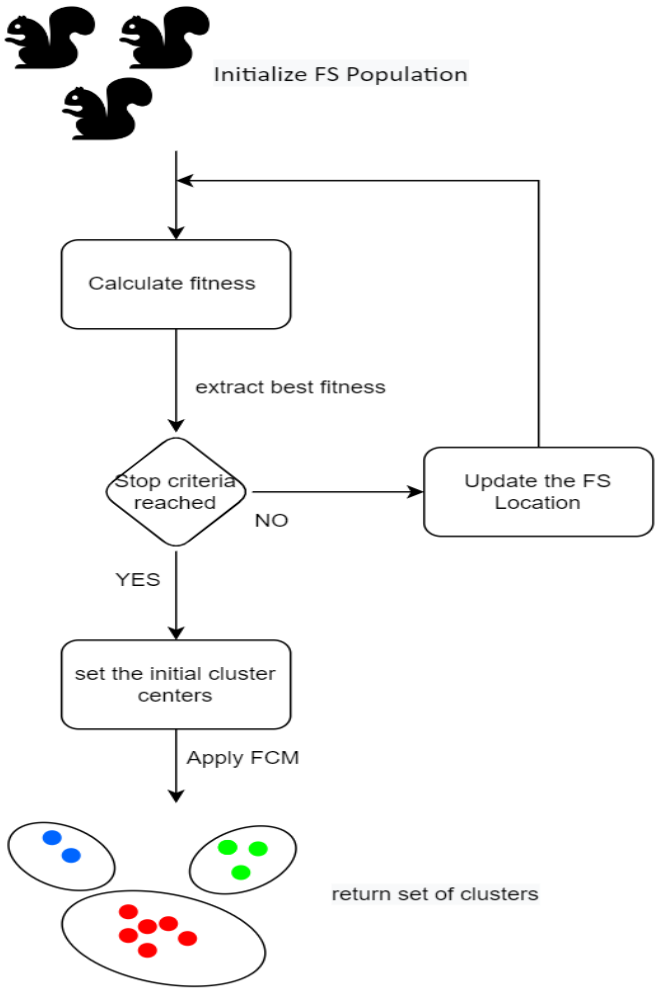

Figure 3 Flowchart of the SSFCA algorithm

6. Expected Research Outcome

1) In this research, an optimized algorithm for vision base page segmentation will be designed and compared its accuracy with traditional vision-based segmentation techniques.
2) In this research, Hybrid algorithms for web page segmentation using clustering and optimizing techniques will be designed, and the hybrid algorithms are compared with the other traditional clustering techniques.
3) In thisresearch,, we will propose two novel approaches for Text Document clustering SSAC and SSFCA as enhancement in the feature of K-means clustering and Fuzzy C-means clustering.
4) Lastly, we will examine the accuracy of SSAC AND SSFCA in view to highlight and enhance the quality of the proposed work.

7. References

1) G.Silambarasan, K.Mohan," Perceiving Syntactic Content Structure of Web Page Using VIPS",*International Journal of Engineering Science and Computing, March 2016*.
2) Prof.K.S.Kadam, Prof.A.B.majgave, Prof.Y.M.Kamble," Extracting Content Structure from Web Pages by Applying Vision based Approach", *International Journal of Computer Science and Information Technologies, Vol. 7 (4) , 2016, 2011-2016.*
3) Vimalkumar B. Vaghela,"Analysis of Various Sentiment Classification Techniques", ,*International Journal of Computer Applications (0975 – 8887)Volume 140 – No.3, April 2016.*
4) Kumud Jaglan," Visual Webpage Content Segmentation and Retrieval based on n-Grams",*Volume 2 Issue – 4Year of Publication : 2015*.
5) K haingwahwah Linn,"Web Page Sefmentation using Combined VIPS",*International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-2, Issue-4, April-2015.*
6) Tingting Wei, Yonghe Lu, Xuanjie Li," Web page segmentation based on the hough transform and vision cues",*Proceedings of APSIPA Annual Summit and Conference 2015*
7) Andr´es Sanoja Vargas,"Web page segmentation, evaluation and applications,Pierre" *Marie Curie - Paris VI, 2015. English. [2015]*.
8) Angelos Valsamis, Alexandros Psychas, Fotis Aisopos, Andreas Menychtas, Theodora Varvarigou, "Lecture Notes in Computer Science", *vol. 10108, pp. 514, 2017, ISSN 0302-9743, ISBN 978-3-319-52835-9[2015]*.
9) Pooja Devi, Ashlesha Gupta, Ashutosh Dixit," Comparative Study of HITS and PageRank Link based Ranking Algorithms",*International Journal of Advanced Research in Computer and Communication EngineeringVol. 3, Issue 2, February 2014 [2014]*.
10) Kuang-Yao Lee,"Visual-Based Web Page Analysis",*San Diego State University [2014]*.
11) P YesuRaju, P KiranSree ,"A Language Independent Web Data Extraction Using Vision Based Page Segmentation Algorithm",*IJRET APR 2013,Volume: 2 Issue: 4,635 - 639,ISSN: 2319 - 1163.*

12) Shefali Singhal," Hybrid Web-page Segmentation and Block Extraction for Small Screen Terminals".*International Journal of Computer Applications (0975 – 8887)4th International IT Summit Confluence 2013 [2013]*.

13) M. Lavanya,M.Dhanalakshmi,"Various approaches of Vision Based Deep Web Data Extraction(VDWDE) and Applications",*Publications of Problem &Application in Engineering Research-Paper[2013]*.

14) M. Elgin Akpınar1 and Yeliz Yes¸ilada," Vision Based Page Segmentation Algorithm: Extended and Perceived Success",*International Conference on Web Engineering [2013]*.

15) P YesuRaju,P Kiran Sree, " A Language Independent Web Data Extraction Using Vision Based Page Segmentation Algorithm.",*IJRET,Volume: 2 Issue: 4,635 - 639,ISSN: 2319 – 1163 APR[2013]*.

16) Deng Cai, Shipeng Yu," A Vision based Page Segmentation Algorithm",*Microsoft Research Lab - Asia [2012]*.

17) Natheer Khasawneh,"Vision-based Presentation Modeling of Web Applications: A Reverse Engineering Approach",,*Journal of Emerging Technologies in Web Intelligence · May [2012]*.

18) Wei Liu, Xiaofeng Meng," A Vision-Based Approach for Deep Web Data Extraction",*IEEE Transactions on Knowledge and Data Engineering, VOL. 22, NO. 3, MARCH 2010 [2010]*

19) Dilip Kumar Sharma," A Comparative Analysis of Web Page Ranking Algorithms", *International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010 [2010]*.

20) Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm",*Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04)0-7695-2096-0/04 $20.00 © 2004 IEEE [2004]*.

21) Wei Liu,Xiaofeng Meng," Vision based Web Data Records Extraction", *Proceedings of 9th Web DBWorkshop International Workshop on Web and Data bases Chicago, Illinois,june 30,2006.*

22) Jan Zeleny,Radek Burget,Jaroslav Zendulka,"Box clustering segmentation :A new method for vision-based web page preprocessing",*Information Processsing & Management,53(3),735-750.*

23) Madhulatha, T. Soni. "An overview on clustering methods." arXiv preprint arXiv:1205.1117 (2012).

24) Gaertler, Marco. "Clustering." In Network analysis, pp. 178-215. Springer, Berlin, Heidelberg, 2005.

25) Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." IEEE Transactions on neural networks 16, no. 3 (2005): 645-678.

26) Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31, no. 3 (1999): 264-323.

27) Von Luxburg, Ulrike, Robert C. Williamson, and Isabelle Guyon. "Clustering: Science or art?." In Proceedings of ICML workshop on unsupervised and transfer learning, pp. 65-79. JMLR Workshop and Conference Proceedings, 2012.

28) Fisher, Lloyd, and JOHN W. VAN NESS. "Admissible clustering procedures." Biometrika 58, no. 1 (1971): 91-104.

29) Reynolds, Alan P., Graeme Richards, Beatriz de la Iglesia, and Victor J. RaywardSmith. "Clustering rules: a comparison of partitioning and hierarchical clustering algorithms." Journal of Mathematical Modelling and Algorithms 5, no. 4 (2006): 475- 504.

30) Barioni, Maria Camila N., Humberto Razente, Alessandra MR Marcelino, Agma JM Traina, and Caetano Traina Jr. "Open issues for partitioning clustering methods: an overview." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4, no. 3 (2014): 161-177

31) Narwal, Neetu, and Sanjay Kumar Sharma. "Machine Learning for Web Page Adpatation." Vol. 7 (1), 2016, 60-65

32) Mohata, P., and Sheetal Dhande. "Web data mining techniques and implementation for handling big data." Computer Science and Mobile Computing 4, no. 4 (2015): 330-334

33) Palekar, V. R., M. S. Ali, and R. Meghe. "Deep web data extraction using webprogramming-language-independent approach." Journal of Data Mining and Knowledge Discovery 3, no. 2 (2012): 69.

34) Liu, Tantan, and Gagan Agrawal. "Stratified k-means clustering over a deep web data source." In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1113-1121. 2012.

35) Wu, Junjie. Advances in K-means clustering: a data mining thinking. Springer Science & Business Media, 2012.

36) Alkhatib, Bassel, and Randa S. Basheer. "Mining the Dark Web: A Novel Approach for Placing a Dark Website under Investigation." International Journal of Modern Education & Computer Science 11, no. 10 (2019).

37) Yang, Bo, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. "Towards k-meansfriendly spaces: Simultaneous deep learning and clustering." In international conference on machine learning, pp. 3861-3870. PMLR, 2017.

38) Zou, Hailei. "Clustering algorithm and its application in data mining." Wireless Personal Communications 110, no. 1 (2020): 21-30

39) Berkhin, Pavel. "A survey of clustering data mining techniques." In Grouping multidimensional data, pp. 25-71. Springer, Berlin, Heidelberg, 2006.

40) Wu, Wensheng, Clement Yu, AnHai Doan, and Weiyi Meng. "An interactive clustering-based approach to integrating source query interfaces on the deep web." In Proceedings of the 2004 ACM SIGMOD international conference on Management of data, pp. 95-106. 2004

41) Kise, Koichi. "Page segmentation techniques in document analysis." In Handbook of Document Image Processing and Recognition, pp. 135-175. Springer, London, 2014.

42) Chaudhary, Meena, and Jyoti Pruthi. "Skip Gram Model Base Entropy Reduction using Vision based Page Segmentation for Web Mining." In 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 842- 846. IEEE, 2019

43) Cormier, Michael. "Computer vision-based analysis of web page structure for assistive interfaces." In Proceedings of the 13th International Web for All Conference, pp. 1-2. 2016

44) Liu, Wei, Xiaofeng Meng, and Weiyi Meng. "Vide: A vision-based approach for deep web data extraction." IEEE transactions on knowledge and data engineering 22, no. 3 (2009): 447-460.

45) Akpınar, M. Elgin, and Yeliz Yesilada. "Vision based page segmentation algorithm: Extended and perceived success." In International Conference on Web Engineering, pp. 238-252. Springer, Cham, 2013.

46) Liu, Wei, and Xiaofeng Meng. "A holistic solution for duplicate entity identification in deep web data integration." In 2010 Sixth International Conference on Semantics, Knowledge and Grids, pp. 267-274. IEEE, 2010.

47) Zhang, Qiang, Xiaoxiao Jiang, and Jiashen Sun. "Web mining based on VIPS in intention-based information retrieval." In 2009 International Conference on Natural Language Processing and Knowledge Engineering, pp. 1-5. IEEE, 2009

48) Sun, Lu, and Jie Zhou. "Development of multiregime speed–density relationships by cluster analysis." Transportation Research Record 1934, no. 1 (2005): 64-71.

MANAV RACHNA
vidyanatariksha

Swami Vivekananda Library and Resource Centre
Manav Rachna University

SIMILARITY VERIFICATION CERTIFICATE

Title: A Hybrid Approach Solution for Web page segmentation Using Clustering and Optimizing Technique.

Researcher: Meene Chaudhary

Registration Number: MRU CSD NO/2K16 0171

Supervisor: Dr. Jyoti Bhatti

Department: C.S.T., MRU

This is to certify that the above mentioned Short Synopsis/Thesis was scanned for similarity index through Turnitin Software on 5-11-2022 with submission ID 1946224604 and the similarity index was found to be 4 %. The complete report running into 28 Pages, 6282 Word count with File Size 3190 KB is being forwarded to the Research Supervisor/DRC/DRB for review.

MANAV RACHNA UNIVERSITY ★ FARIDABAD ★

5/11/2022

Date: 5/11/2022

The complete report of the above Short Synopsis/Thesis has been reviewed and the similarity index is below the accepted norms as per the University Grants Commission (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations 2018. The Short Synopsis/Thesis may be considered for further processing

Meene Chaudhary
Research Scholar

Jyoti
Research Supervisor

A Hybrid Approach Solution for web page segmentation using Clustering and Optimizing Techniques

by Meena Chaudhary

mission date: 05-Nov-2022 01:11PM (UTC+0530)
mission ID: 1945224604
name: page_segmentation_using_Clustering_and_Optimizing_Techniques.doc (404K)
rd count: 6282
racter count: 35256

A Hybrid Approach Solution for web page segmentation using Clustering and Optimizing Techniques

SHORT SYNOPSIS

by

Meena Chaudhary
2K16-PFE-012

Under the supervision of

Dr. Jyoti Pruthi
Professor

Department of Computer Science and Technology
School of Engineering

MANAV RACHNA Vidyantarikshal | MANAV RACHNA UNIVERSITY
Declared as State Private University vide Haryana Act 26 of 2014

Year *2018*

ABSTRACT

Data extraction from the Web is a popular research topic as of late, with most suggested methods depending on parsing HTML code or tag trees. More and more Web databases are being available online and may be queried using Web interfaces. Traditional crawler-based search engines have a hard time indexing these unique Web sites because of the way they are created. The following are the primary restrictions of these approaches: To begin with, they rely on HTML, the language used to create Web pages. However, HTML is constantly evolving. So as new versions and tags are released, older works will need to be revised accordingly. The question raised by previous answers is whether major customization or full assignment is required. Second, it cannot handle the increasingly complex HTML used on websites. Website designers employ more sophisticated implementations of JavaScript and CSS to create visually striking and engaging websites. Compare different websites based on embedded text, images, video files, hyperlinks, etc. using Vision-based Page Segmentation

Algorithm (VIPS). Gram models and lemmatization capabilities for implementing deep web data extraction from unstructured sources.

The Internet and modern applications include vast amounts of material, which may be overwhelming to sort through. Organizing massive amounts of text documents into meaningful clusters, as the text clustering approach does, is a useful tool for handling such data. Larger documents have less of an impact on the efficiency of the text clustering method. Accordingly, the efficacy of the text clustering approach is diminished due to the presence of sparse and uninformative features (i.e., noisy, irrelevant, and superfluous characteristics) in the text documents themselves. Meta-heuristic algorithms have recently been used to effectively address various challenging optimization issues. This thesis highlights the aforementioned clustering issue and suggests a few metaheuristic methods for resolving it.

Key words: skip gram, VIPS - Vision based Page Segmentation, Deep Web, *Text Clustering, Segmentation*

A Hybrid Approach Solution for web page segmentation using Clustering and Optimizing Techniques

ORIGINALITY REPORT

4%	2%	4%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

Excluding matches less than 14 words

PRIMARY SOURCES

1	"International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2019 Publication	1%
2	bioinfopublication.org Internet Source	1%
3	docplayer.net Internet Source	1%
4	Yue Gao, Xiangling Fu, Xien Liu, Ji Wu. "Chapter 38 Multi-features-Based Automatic Clinical Coding for Chinese ICD-9-CM-3", Springer Science and Business Media LLC, 2021 Publication	1%
5	Pusdekar, Satish J., and Shaikh Phiroj Chhaware. "Using Visual Clues Concept for Extracting Main Data from Deep Web Pages", 2014 International Conference on Electronic Systems Signal Processing and Computing Technologies, 2014.	<1%

Publication

6	link.springer.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		